ACADEMIC PRESS

Available at WWW.MATHEMATICSWEB.ORG POWERED BY SCIENCE DIRECT®

Journal of Multivariate Analysis 86 (2003) 183–212

Journal of Multivariate Analysis

http://www.elsevier.com/locate/jmva

# Clusters, outliers, and regression: fixed point clusters

Christian Hennig[a,b,*]

[a] *Fachbereich Mathematik - SPST, Universität Hamburg, Bundesstraße 55, D-20146 Hamburg, Germany*
[b] *Seminar für Statistik, ETH Zentrum, CH-8092 Zürich, Switzerland*

Received 21 May 2001

**Abstract**

Fixed point clustering is a new stochastic approach to cluster analysis. The definition of a single fixed point cluster (FPC) is based on a simple parametric model, but there is no parametric assumption for the whole dataset as opposed to mixture modeling and other approaches. An FPC is defined as a data subset that is exactly the set of non-outliers with respect to its own parameter estimators. This paper concentrates upon the theoretical foundation of FPC analysis as a method for clusterwise linear regression, i.e., the single clusters are modeled as linear regressions with normal errors. In this setup, fixed point clustering is based on an iteratively reweighted estimation with zero weight for all outliers. FPCs are non-hierarchical, but they may overlap and include each other. A specification of the number of clusters is not needed. Consistency results are given for certain mixture models of interest in cluster analysis. Convergence of a fixed point algorithm is shown. Application to a real dataset shows that fixed point clustering can highlight some other interesting features of datasets compared to maximum likelihood methods in the presence of deviations from the usual assumptions of model based cluster analysis.


*AMS 1991 subject classifications:* 62H30; 62J05; 62-07

*Keywords:* Fixed point clusters; Clusterwise linear regression; Mixture model; Outlier identification; Redescending M-estimators

*Corresponding author. Seminar für Statistik, ETH Zentrum, CH-8092 Zürich, Switzerland.
*E-mail address:* hennig@math.uni-hamburg.de, hennig@stat.math.ethz.ch.

0047-259X/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0047-259X(02)00020-9

## 1. Introduction

Cluster analysis is related to the concept of outliers. If a part of a dataset forms a well-separated cluster, this means that the other points of the dataset appear outlying with respect to the cluster. It may be interpreted synonymously that "a cluster is homogeneous" and that "it does not contain any outlier". The idea of fixed point clusters (FPCs) is to formalize a cluster as a data subset that does not contain any outlier and with respect to which all other data points are outliers. It is rooted in robust statistics as explained in Section 2.

The concept is applied to clusterwise linear regression in this paper. That is, a relation

$$y = \mathbf{x}'\boldsymbol{\beta} + u, \quad E(u) = 0,$$

between a dependent variable $y$ and an independent variable $\mathbf{x} \in \mathbb{R}^p \times \{1\}$ ($\beta_{p+1}$ denoting the intercept parameter) should be adequate for a single cluster. Fig. 1 shows data from the Old Faithful Geyser in the Yellowstone National Park, collected in August 1985. The duration of an eruption of the geyser is modeled here as dependent on the waiting time since the previous eruption. One can recognize

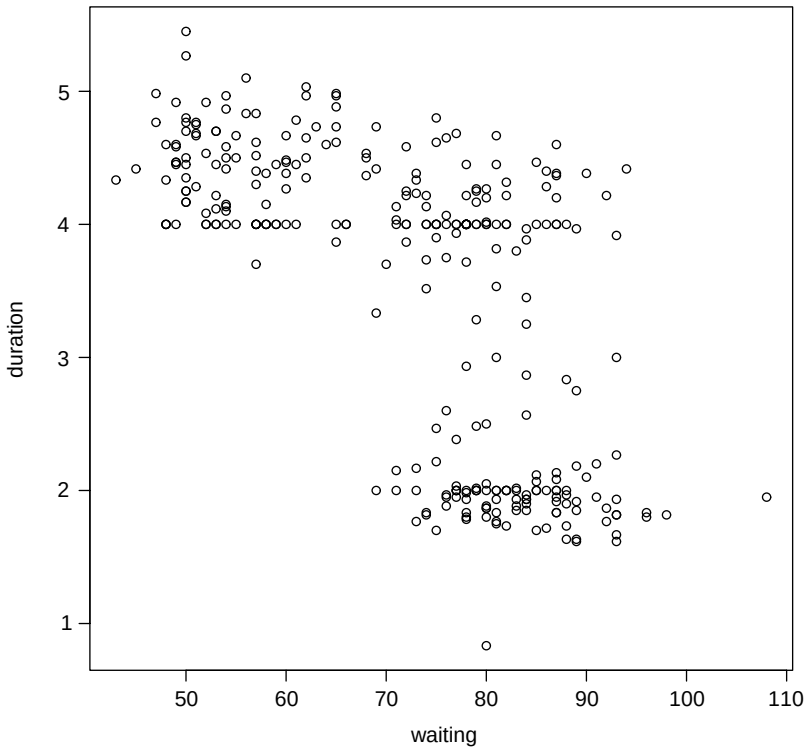


Fig. 1. Old Faithful Geyser data.

roughly two groups of linear dependence between "waiting" and "duration", corresponding to the eruptions with lower and higher duration, besides other features, which are discussed in more detail in Section 8. The data were taken from Azzalini and Bowman [2]. The aim of clusterwise linear regression is to find such kind of heterogeneity. Further applications of clusterwise linear regression appear e.g. in biology [21] and market segmentation [12,29].

The literature on clusterwise linear regression concentrates mainly on least squares and maximum likelihood methods for mixture and partition models [13,21] more references are given in [12]. A method for mixtures of multivariate normal distributions was proposed as well for "linearly shaped clusters" [9]. In that case, the independent variable is assumed to be normally distributed.

FPC analysis differs from these concepts, because their models assume that the whole dataset consists of clusters of the same parametric form, namely linear regressions or multivariate normal distributions. In contrast, FPC analysis searches for a single cluster at a time. The only assumption for the rest of the data is that it has to consist of outliers with respect to the cluster. Consequently, not every data point needs to be included in an FPC, and FPCs may intersect. The data in Fig. 1 illustrate the usefulness of these properties: The higher duration points, which as a whole can be interpreted as a cluster, contain some points exactly on a line, and some of the data points do not fit in any visible cluster.

It has to be noted that methods based on mixture models can cope as well with such features: A posteriori membership probabilities can be considered to interpret clusters as intersecting, although these probabilities have to sum up to 1, and usually the points are assigned only to the cluster where they have the largest membership probability. DasGupta and Raftery [9] add a Poisson process component to their normal mixture to capture outlying points. But this adds a parametric assumption for the non-normal part of the data.

Furthermore, FPCs by definition remain unaffected under changes and distortions in the region of outlying points, and this cannot hold for estimation methods that fit the whole dataset simultaneously. FPC analysis is not meant as a new method to estimate mixture components. It provides an alternative *definition* of a cluster [18].

Apart from the mixture paradigm, Cook and Critchley [5] develop a method to detect linear regression clusters graphically. Morgenthaler [25] uses the local minima of redescending M-estimators to find such clusters. FPC analysis can be considered as a generalization of his approach, see Section 2.

For the sake of simplicity, I discuss the one-dimensional location clustering problem (i.e., linear regression without slope, $p = 0$) to motivate the idea of FPC analysis in Section 2. Least squares-FPCs for the linear regression setup are defined in Section 3. Section 3.2 introduces a convergent algorithm to find FPCs. Section 4 discusses generalizations of the approach. Conditions for the consistency of least squares-FPCs for theoretical LS-FPCs, i.e., FPCs defined for distributions, are given in Section 5. Theoretical FPCs are calculated for certain mixture distributions and the consistency conditions are checked in Section 6. Section 7 describes an implementation of the method along with the choice of all required constants. In Section 8, FPC analysis is applied to the Old Faithful Geyser data and compared to

the results of the mixture model estimators of DeSarbo and Cron [12] and Fraley and Raftery [14].

A short conclusion is given in Section 9. Some lemmas and theorems are proven in the appendix. All detailed proofs are given in [20].

Here is some notation. "$||\mathbf{x}||$" denotes the Euclidean norm of $\mathbf{x}\in\mathbb{R}^q$. $\mathbf{I}_p$ denotes the $p\times p$-unit matrix. For $\varepsilon>0$, $B_\varepsilon(\mathbf{x})$ is the closed $\varepsilon$-ball around $\mathbf{x}$ w.r.t. the Euclidean metric. For a given probability distribution $P$, $k\in\mathbb{N}\cup\{\infty\}$, $P^k$ symbolizes the $k$-fold independent product. $P^\infty$ is used as the parent distribution for i.i.d. random variables $z_1,\ldots,z_n$ with $\mathscr{L}(z_1)=P$, which means that $z_1$ is distributed according to $P$. $P_k$ stands for the empirical distribution according to $(z_1,\ldots,z_k)$ where $\mathscr{L}(z_1,\ldots,z_k)=P^k$. I write $Pf$ for $\int f\,\mathrm{d}P$. $1[z\in B]$ denotes the indicator function of the set $B$.

## 2. Clusters, outliers, $M$-estimators and fixed points

The link between outlier identification, robust statistics and cluster analysis is mentioned first by Hampel et al. [16, p. 46], to my knowledge. Robust statistics often deals with the location of a large homogeneous "main part" of the data in presence of outliers, which may be produced by mechanisms different from sources of the rest, and which should not largely affect the estimation of the main part. Cluster analysis more generally aims to locate any homogeneous part of the data. The recognition of such a part should not be strongly affected by changes in distant parts of the data. This demand is violated by many CA methods, in particular by partitioning methods such as $k$-means (see [15]). If there is a clear separation between main part and outliers, the main part can be regarded as the largest cluster, and robust statistics may serve to find it. But it can also point to the other ones, as explained in the following.

Imagine a one-dimensional dataset $(y_1,\ldots,y_n)$, $n=30$, with 20 observations from $\mathcal{N}(0,1)$ (avoiding the extreme tail areas), 5 observations from $\mathcal{N}(10,1)$, and 5 observations from $\mathcal{N}(30,1)$, i.e., three strongly separated clusters. M-estimators $T_\rho$ of location (see e.g. [22]) are defined by solving

$$\sum_{i=1}^{n}\rho\left(\frac{y_i-T_\rho}{s}\right)=\min \tag{1}$$

with a suitable chosen loss function $\rho$ and a scale $s>0$, or alternatively by

$$\sum_{i=1}^{n}\psi\left(\frac{y_i-T_\rho}{s}\right)=0, \tag{2}$$

where $\psi=\rho'$ (possibly piecewise). A solution of (2) is a fixed point of

$$f(t):=\frac{\sum_{i=1}^{n}w((y_i-t)/s)y_i}{\sum_{i=1}^{n}w((y_i-t)/s)},\quad w(z):=\frac{\psi(z)}{z}. \tag{3}$$

That is, $T_\rho$ is a weighted mean, where the weights depend on $T_\rho$ itself. It may be obtained by the ordinary fixed point algorithm under certain conditions [22, p. 146]. In linear regression such algorithms are sometimes called "iteratively reweighted least squares" [25]. $w((y_i - t)/s)$ gives the weight of $y_i$ for the computation of $t$ and may be interpreted as a measure of centrality (outlyingness, respectively) of the point $y_i$ with respect to $t$.

For example, the median corresponds to $\rho(z) = z*1[z>0] - z*1[z<0]$ regardless of $s$. As many robust location estimators, it will appear close to 0 for the data above, but positively biased (if interpreted as estimator for the data from $\mathcal{N}(0,1)$) because of the asymmetrical contamination in positive direction.

The bias may be avoided by the so-called "redescending M-estimators", which are M-estimators with $\rho(z)$ constant for large absolute values of $z$, and therefore $\psi(z) = w(z) = 0$. Such points do not have any weight for the computation of $T_\rho$, as desired for outliers. If $s$ is chosen small enough, such an estimator estimates the center of $\mathcal{N}(0,1)$ unaffected by any point from the smaller populations. Furthermore, such $T_\rho$ remains a solution of (3) under addition or deletion of outliers in the sense of this definition, i.e., of points with $w((y - T_\rho)/s) = 0$. But a solution of (2), (3), respectively, is usually not unique for redescending M-estimators. If $s$ is chosen such that $w((y - t)/s) = 0$ holds for $|y - t| > 4$, say, there will be solutions estimating the centers of $\mathcal{N}(10,1)$ and $\mathcal{N}(30,1)$ as well, since the "window" of points with positive weight $w$ around the center of each of the three clusters will only contain points from the same cluster. This leads to the thought that the solutions of (3) for redescending M-estimators might be used to locate an unknown number of clusters stably in the presence of outliers.

The main problem is the choice of $s$. In robust statistics one often uses a preliminary robust estimate of scale, for example the MAD. But such an estimate depends on at least half of the points. That is, if the largest cluster contains fewer than half of the points, $s$ depends upon points of at least two clusters and gets too large for a single cluster. Furthermore, the clusters may have differing scales. If $\mathcal{N}(30,1)$ would be replaced by $\mathcal{N}(30,6)$, a weight window adjusted to variance 1 may capture only few points of this component, while working with variance 6 may destroy the separation between the other two populations.

The idea of FPC analysis is to define the location (regression parameters, respectively) and scale estimators jointly via a fixed point condition using only the corresponding non-outliers, so that both parameters are adapted to the local cluster. Such parameter estimators can no longer be described as minima of some global criterion like (1), since there is no natural ordering of quality among them. The weights will be chosen so that they can only take the values 0 (outlier) and 1 (non-outlier). That is, a solution of (3) is characterized as corresponding to a subpopulation (defined by the weights for all points) that is exactly the set of non-outliers w.r.t. its own parameter estimators. This corresponds to a $\psi$-function as shown in [16, p. 159]. A generalization to continuous choices of $w$, leading to fuzzy clusterings, is possible.

The resulting estimators fall into the class of simultaneous M-estimators of location and scale as defined by [22, p. 136], but the theory given there does exclude redescending $\psi$-functions. Adrover and Yohai [1] establish asymptotic normality for a class of simultaneous redescending M-estimates of regression and scale, which are defined in order to *avoid* the occurrence of solutions belonging to differing subpopulations. To my knowledge, Morgenthaler [25] was the first author to investigate the use of redescending M-estimators for the location of groups and multiple patterns of the data. He discussed the choice of $s$ in a linear regression setup based on the MAD of residuals of the LS-estimator as well as using a decreasing sequence of values for $s$, but he did not treat clusters with differing scales. Similar ideas are known in image analysis, see e.g. [4].

Some referees of a previous version of this paper wondered why redescending M-estimators are used for FPC analysis instead of high breakdown methods as discussed by Rousseeuw and Leroy [27] or Rousseeuw and Hubert [26]. Such methods are able to cope with clustered outliers, and it seems to be natural to connect them to cluster analysis. But all these estimators search for optimal parameter values for a majority of the data. They assume that more than half of the data points belong to a homogeneous population. A breakdown point of 50% makes no sense if the goal is to find parameters for several clusters, none of which needs to contain half of the data. The advantage of redescending M-estimators in this setup is that they give zero weight to all the outliers with respect to a given subset of the data regardless of the number of members of the subset.

Alternative suggestions for the use of robust techniques in cluster analysis were made by Davies [10] and Cuesta-Albertos et al. [8]. Several authors tried to robustify standard methods for mixture models by the insertion of robust estimators, including redescending M-estimators. For references see Section 7 of McLachlan and Peel [24]. All these approaches belong to the mixture or partition paradigm and are not generalized to linear regression clusters up to now.

## 3. Fixed point clusters in linear regression

### 3.1. Definition for datasets

Let $\mathbf{Z} := \mathbf{Z}_n := (\mathbf{X}, \mathbf{y}) := ((\mathbf{x}_1', y_1), \ldots, (\mathbf{x}_n', y_n))'$, where $\mathbf{x}_i \in \mathbb{R}^p \times \{1\}$, $y_i \in \mathbb{R}$, $i = 1, \ldots, n$, be a regression dataset. For a given indicator (weight) vector $\mathbf{w} \in \{0,1\}^n$ let $\mathbf{Z}(\mathbf{w}) = (\mathbf{X}(\mathbf{w}), \mathbf{y}(\mathbf{w}))$ be the dataset consisting only of the points $(\mathbf{x}_i', y_i)$ with $w_i = 1$. $n(\mathbf{w})$ represents the number of points indicated by $\mathbf{w}$. For FPC analysis in the regression setup, particular weight vectors are of interest. They indicate the points lying close to the regression hyperplane defined by a parameter $\boldsymbol{\beta}$ in terms of a variance parameter $\sigma^2$:

$$\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2} := (1[(y_i - \mathbf{x}_i'\boldsymbol{\beta})^2 \leqslant c\sigma^2])_{i=1,\ldots,n}.$$

An FPC is a data subset defined by some weight vector $\mathbf{w}$ indicating the non-outliers w.r.t. to the LS-estimator $\hat{\beta}(\mathbf{Z}(\mathbf{w}))$ weighted by $\mathbf{w}$ itself. Outlyingness is measured by means of the weighted error variance estimator $\hat{\sigma}(\mathbf{Z}(\mathbf{w}))$. These are parameter estimators satisfying a fixed point condition analogously to (3). A tuning constant $c>1$ has to be chosen to define the tolerance of the outlier classification (see Sections 4 and 7).

**Definition 3.1.** An indicator vector $\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2} \in \{0,1\}^n$ is called *least squares-fixed point cluster vector* (LS-FPCV) w.r.t. $\mathbf{Z}$ (and the indicated points form an LS-FPC), iff $(\boldsymbol{\beta},\sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ is a fixed point of

$$f_{\mathbf{Z}} : (\boldsymbol{\beta},\sigma^2) \mapsto (\hat{\beta}[\mathbf{Z}(\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2})], \hat{\sigma}^2[\mathbf{Z}(\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2})]),$$

where

$$\hat{\beta}(\mathbf{Z}(\mathbf{w})) := (\mathbf{X}(\mathbf{w})'\mathbf{X}(\mathbf{w}))^{-1}\mathbf{X}(\mathbf{w})'\mathbf{y}(\mathbf{w}),$$

$$\hat{\sigma}^2(\mathbf{Z}(\mathbf{w})) := \frac{1}{n(\mathbf{w})-p-1}\sum_{i=1}^{n} w_i(y_i - \mathbf{x}_i'\hat{\beta}(\mathbf{Z}(\mathbf{w})))^2.$$

In case of the non-existence of $(\mathbf{X}(\mathbf{w})'\mathbf{X}(\mathbf{w}))^{-1}$, $f_{\mathbf{Z}}(\boldsymbol{\beta},\sigma^2) := (\boldsymbol{\beta},\infty)$.

For example, consider the points indicated by triangles in Fig. 2. They are indicated by the weight vector $\mathbf{w} = \mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2}$, where $\boldsymbol{\beta}$ corresponds to the solid line and the dotted lines show $\mathbf{x}'\boldsymbol{\beta} \pm \sqrt{c\sigma^2}$. They form an LS-FPC for $c = 6.635$, since one finds $(\hat{\beta}(\mathbf{Z}(\mathbf{w})), \hat{\sigma}^2(\mathbf{Z}(\mathbf{w})) = (\boldsymbol{\beta},\sigma^2)$.

Consider on the other hand the squares with values of "duration" between 2 and 4. If the LS-regression line is estimated for these data, their error variance is so large that some of the circles and some of the triangles would get inside the corresponding strip. This would make the error variance of the resulting data subset even larger and it would also change the regression line, so that the fixed point condition is not fulfilled and this data subset is not separated enough from the rest to form an LS-FPC. The full result for the Geyser data is discussed in Section 8.

Note that FPCs may intersect or include each other. In particular, all subsets $\mathbf{Z}(\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2})$ with $\sigma^2 = 0$ and non-collinear covariate points form LS-FPCs. The implementation described in Section 7 avoids to find trivial meaningless LS-FPCs.

Since the FPC-property of a subset does only depend upon the points inside the strip defined by its parameter estimators, the deletion of any of the points outside the three FPCs of Fig. 2 (i.e., the points denoted by squares), or the addition of such points, would not change the FPC-property of any of these clusters.

### 3.2. A fixed point algorithm for LS-fixed point cluster vectors

It is practically impossible to check the FPC-property of every subset of a dataset, except if it is very small. But LS-FPCVs can be found by means of a fixed point algorithm. Theorem 3.1 guarantees its convergence. The use of the algorithm for an

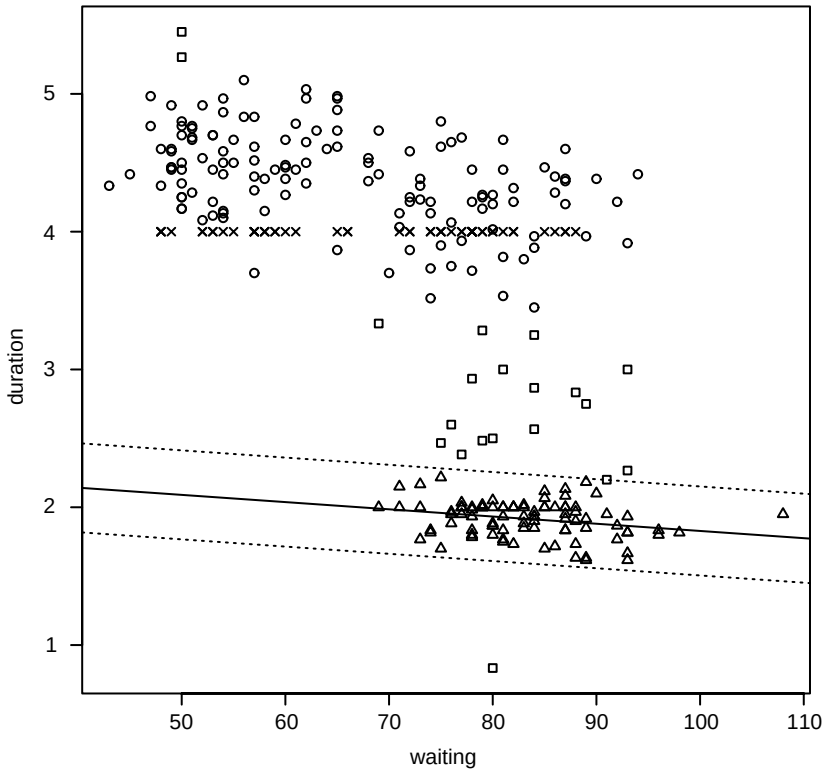


Fig. 2. Old Faithful Geyser data with LS-FPCs, $c = 6.635$. The points indicated by crosses form an FPC as well as the triangles. A further FPC consists of the circles together with the crosses.

implementation of FPC analysis is described in Section 7. The convergence result is needed for the consistency theory of Section 5 as well.

**Fixed point algorithm (FPA):.** Choose $\mathbf{w}^0 \in \{0,1\}^n$ with $n(\mathbf{w}^0) > p+1$, $k=0$.
*Step* 1: Compute $\hat{\beta}(\mathbf{Z}(\mathbf{w}^k))$, $\hat{\sigma}^2(\mathbf{Z}(\mathbf{w}^k))$.
*Step* 2: $\mathbf{w}_i^{k+1} = \mathbf{w}_{\mathbf{Z}i}(\mathbf{w}^k) := 1((y_i - \mathbf{x}_i'\hat{\beta}(\mathbf{Z}(\mathbf{w}^k)))^2 \leqslant c\sigma^2(\mathbf{Z}(\mathbf{w}^k)))$, $i = 1, \ldots, n$.
*Step* 3: End if $\mathbf{w}^k = \mathbf{w}^{k+1}$, else $k = k+1$, step 1.

**Theorem 3.1.** *Let* $c > 1$. *If* $(\mathbf{X}(\mathbf{w})'\mathbf{X}(\mathbf{w}))^{-1}$ *exists for all* $\mathbf{w} \in \{0,1\}^n$ *with* $n(\mathbf{w}) > p+1$, *then for some* $k < \infty$ : $\mathbf{w}^k = \mathbf{w}_{\mathbf{Z}}(\mathbf{w}^k)$, *i.e., the FPA converges in finitely many steps.*

The proof is given in [30].

### 3.3. Definition for distributions

In order to investigate the statistical properties of LS-FPC analysis, I define a distribution version of LS-FPCs. Let $P$ denote a distribution on $\mathbb{R}^p \times \{1\} \times \mathbb{R}$, i.e., a

distribution for regression data points $(\mathbf{x}', y)$ as above. LS-FPCs of a distribution should consist of all points of appropriate strips around regression hyperplanes where the distribution is "regression cluster-shaped". They are indicated by weight functions of the form

$$w_{\boldsymbol{\beta},\sigma^2}(\mathbf{x}, y) := 1[(y - \mathbf{x}'\boldsymbol{\beta})^2 \leqslant c\sigma^2].$$

For a measurable indicator function $w$ let $P_w$ be the conditional distribution of $P$ under $\{w = 1\}$, i.e., the restriction of $P$ to the points indicated by $w$.

Theoretical LS-FPCs of distributions are defined by replacement of the regression and scale estimators by their corresponding functionals in the definition of LS-FPCVs.

**Definition 3.2.** An indicator function $w_{\boldsymbol{\beta},\sigma^2}$, $(\boldsymbol{\beta}, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}_0^+$ is called *least squares-fixed point cluster indicator* (LS-FPCI) w.r.t. $P$, iff $(\boldsymbol{\beta}, \sigma^2)$ is a fixed point of

$$f_P : (\boldsymbol{\beta}, \sigma^2) \mapsto (\tilde{\beta}[P_{w_{\boldsymbol{\beta},\sigma^2}}], \tilde{\sigma}^2[P_{w_{\boldsymbol{\beta},\sigma^2}}]),$$

where

$$\tilde{\beta}(P_w) := \arg\min_{\beta} P_w(y - \mathbf{x}'\boldsymbol{\beta})^2,$$
$$\tilde{\sigma}^2(P_w) := P_w(y - \mathbf{x}'\tilde{\beta}(P_w))^2.$$

If $\arg\min_\beta P_w(y - \mathbf{x}'\boldsymbol{\beta})^2$ is not defined uniquely, $f_P(\boldsymbol{\beta}, \sigma^2) := (\boldsymbol{\beta}, \infty)$.

The latter implies $Pw_{\boldsymbol{\beta},\sigma^2} > 0$ for all LS-FPCIs. (4)

Under suitable conditions, LS-FPCVs turn out to be consistent estimators for LS-FPCIs in Section 5. That is, LS-FPCVs can be viewed as estimators of clusters of distributions, if the LS-FPCIs indicate such plausible clusters. This is discussed in Section 6.

The components of the functions $f_{\mathbf{Z}}, f_P$, respectively, are written as follows from now on: $\beta_{\mathbf{Z}}(\boldsymbol{\beta}, \sigma^2) := \hat{\beta}[\mathbf{Z}(\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2})]$, $\sigma^2_{\mathbf{Z}}(\boldsymbol{\beta}, \sigma^2) := \hat{\sigma}^2[\mathbf{Z}(\mathbf{w}_{\mathbf{Z},\boldsymbol{\beta},\sigma^2})]$, $\beta_P(\boldsymbol{\beta}, \sigma^2) := \tilde{\beta}[P_{w_{\boldsymbol{\beta},\sigma^2}}]$, $\sigma^2_P(\boldsymbol{\beta}, \sigma^2) := \tilde{\sigma}^2[P_{w_{\boldsymbol{\beta},\sigma^2}}]$.

**Remark 1.** The regression equivariance properties of the LS- and variance estimators carry over to FPCVs and FPCIs, i.e., $w_{\boldsymbol{\beta},\sigma^2}$ is an LS-FPCI w.r.t. $P$ iff $w_{\boldsymbol{\beta},\sigma^2} \circ D = w_{(\mathbf{A}^{-1})'(a\boldsymbol{\beta}+b),a^2\sigma^2}$ is an LS-FPCI w.r.t. $P^D$ under linear transformations of the form

$$D : \mathbb{R}^{p+2} \mapsto \mathbb{R}^{p+2}, \quad (\mathbf{x}, y) \mapsto (\mathbf{A}\mathbf{x}, ay + \mathbf{x}'b),$$

$\mathbf{A} \in \mathbb{R}^{(p+1)^2}$ invertible with $(0, \ldots, 0, 1)$ as last column, $a \in \mathbb{R}\backslash\{0\}$, $b \in \mathbb{R}^{p+1}$. This holds analogously for FPCVs. The proof is straightforward.

## 4. Fixed point clusters—general

Here is a rougher description of FPCs: Consider a subset of a dataset. Decide for all points of the data subset, whether they are close to the subset (represented by its regression and scale parameter estimator) or lie out. If the non-outlying points are exactly the points of the subset, the subset forms an FPC. That is, the FPC property defines *homogeneity* (no outlier included) and *separateness* (all others are outliers) of a cluster in terms of outlier identification.

This description may be generalized to arbitrary clustering problems. Only an outlier identification rule is needed, that divides the whole dataset into outliers and non-outliers w.r.t. any given subset. The subsets, which do not contain any outlier, and w.r.t. which the whole rest of the data consists of outliers, are the FPCs. An application to multivariate normal clustering is given in [17].

Appropriate outlier identifiers can be found as follows: Davies and Gather [11] emphasized that a definition of the term "outlier" should rely on the idea of an underlying distribution of the homogeneous part of the data. They define "outlier regions" (ORs) as atypical regions of such "reference distributions". For example, in the linear regression case the class of distributions of the type $P_{\boldsymbol{\beta},\sigma^2,G}$ can be considered as the class of reference distribution for homogeneous data, where $P_{\boldsymbol{\beta},\sigma^2,G}$ is defined as the common distribution of $(\mathbf{x}, y)$ according to

$$y = \mathbf{x}'\boldsymbol{\beta} + u, \quad \mathscr{L}(u) = \mathscr{N}(0, \sigma^2), \quad \mathscr{L}(\mathbf{x}) = G, \tag{5}$$

i.e., a model with random covariates, where

$\mathbf{x}$ and $u$ are stochastically independent,

$G$ is any distribution fulfilling

$$G||\mathbf{x}||^2 < \infty, \quad G\mathbf{x}\mathbf{x}' \text{ invertible.} \tag{6}$$

Then,

$$A(\alpha, P_{\boldsymbol{\beta},\sigma^2,G}) := \{(\mathbf{x}, y) \in \mathbb{R}^{p+1} : (y - \mathbf{x}'\boldsymbol{\beta})^2 > c\sigma^2\},$$

$c := c(\alpha)$ being the $(1-\alpha)$-quantile of the $\chi^2_1$-distribution, defines an $\alpha$-OR in the sense of Davies and Gather, i.e., $A(\alpha, P_{\boldsymbol{\beta},\sigma^2,G}) = \alpha$ so that the points in the area of low density of the error distribution are defined as outliers. For example, $c(0.01) = 6.635$. In the definition of LS-FPCs, the parameters $\boldsymbol{\beta}$ and $\sigma^2$ are simply replaced by estimators.

That is, an OR is estimated on the basis of the data subset under consideration. This subset is treated as a set of non-outliers coming from a member of the family of reference distributions, and the whole dataset is treated as generated by a distribution of the form

$$(1-\varepsilon)P_0 + \varepsilon P^*, \quad 0 \leqslant \varepsilon < 1, \tag{7}$$

where $P_0$ is a reference distribution for homogeneous data, and $P^*$ is arbitrary, but should be concentrated on $A(\alpha, P_0)$ with appropriate $\alpha$. Models of the form (7) are called "contamination models". They are often used in robust statistics (e.g. [22]).

Mixture models of the form

$$\sum_{i=1}^{k} \varepsilon_i P_i, \quad \sum_{i=1}^{k} \varepsilon_i = 1, \quad \varepsilon_i > 0, \ i = 1, \ldots, k \tag{8}$$

are more familiar in cluster analysis (e.g. [12]), where $P_i$, $i = 1, \ldots, k$ are cluster reference distributions with distinct parameters. They are of the contamination type (7) as well, but they assume a particular structure for $P^*$, namely, being a mixture of further reference distributions.

From the viewpoint of robust outlier identification, it is questionable to estimate an OR by use of non-robust estimators like the LS-regression estimator. If a dataset (or a subset) contains outliers, they will affect such estimators. Davies and Gather [11] discuss alternative outlier identifiers for the case $p = 0$ and show the superiority of identifiers based on robust estimators for the problem of finding large outliers in the presence of multiple outliers. FPCs may be defined by the use of more general estimators of ORs. The most obvious idea is the replacement of regression and scale parameters $\hat{\beta}$ and $\hat{\sigma}^2$ of the definition of LS-FPCVs, the corresponding functionals of the definition of LS-FPCIs, respectively, by more robust alternatives.

I concentrate on the LS-version here for reasons of computational and theoretical simplicity. Its non-robustness may do less damage for the purposes of cluster analysis, since the aim is to find outlier-free data subsets, and there is no robustness problem for the data subsets which are *in fact* homogeneous and well separated. Recall from Section 2 that LS-FPCVs are based on *redescending* M-estimators as opposed to an LS-estimator for the whole dataset.

For heterogeneous data subsets, however, the estimated OR may get very large, so that there is usually an additional FPC corresponding to (almost) the whole dataset, even if the latter consists of some clearly separated clusters. This is illustrated in the example at the end of Section 6.

## 5. Consistency of LS-fixed point cluster vectors

The LS-FPCIs of the models are the "theoretical clusters" to be estimated by the LS-FPCVs. FPC analysis is intended to be a reasonable tool to analyze data from contamination models (7) where the component $P_0 = P_{\boldsymbol{\beta}_1,\sigma_1,G}$ is well separated from $P^*$. Therefore it is desirable that the parameters of the LS-FPCVs are consistent for the parameters $(\boldsymbol{\beta}_1, \sigma_1^2)$ in some sense. Here are some aspects of the consistency of FPCs:

1. Do LS-FPCVs estimate LS-FPCIs consistently?
   (i) If $P$ has an LS-FPCI, there should be a sequence of LS-FPCVs which is consistent for it (Theorem 5.1).
   (ii) For large enough $n$, all LS-FPCVs should appear close to some LS-FPCI of $P$ with large probability (Corollary 1). Note that there is no result relating the *number* of LS-FPCVs to that of LS-FPCIs.)

2. Do LS-FPCIs adequately reflect the structure of distributions of the contamination type (7)?
   (i) The contamination model should have an LS-FPCI belonging to $P_{\beta_1,\sigma_1,G}$, if it is well separated from $P^*$ (Theorem 6.1, Corollary 2, Fig. 3).
   (ii) $P^*$ may contain further parts of the type $P_{\beta,\sigma^2,G}$. Therefore, it is not reasonable to expect that the LS-FPCI mentioned above would be the only one. But $P$ should not have LS-FPCIs in areas where it does not give rise to any clustering of the data (Theorem 6.1, Fig. 3).
   (iii) If the LS-FPCIs correspond to well separated components of the type $P_{\beta,\sigma^2,G}$, they should fulfill the assumptions of the consistency results (Lemma 6.1, Fig. 3).

In the following, $P$ denotes a distribution on $\mathbb{R}^p \times \{1\} \times \mathbb{R}$, where $\mathscr{L}(\mathbf{Z}_n) = P^n$, $n \in \mathbb{N}$.

The basic result for the asymptotic existence of LS-FPCVs close to the LS-FPCIs, and the non-existence elsewhere, respectively, is the uniform consistency of $f_{\mathbf{Z}_n}(\boldsymbol{\beta}, \sigma^2)$ for $f_P(\boldsymbol{\beta}, \sigma^2)$ for all $(\boldsymbol{\beta}, \sigma^2)$ belonging to a suitable set.

Let $C$ be a compact subset of $\mathbb{R}^{p+1} \times \mathbb{R}_0^+$. Define

$$V(C) := \bigcup_{(\boldsymbol{\beta},\sigma^2) \in C} \{(y - \mathbf{x}'\boldsymbol{\beta})^2 \leqslant c\sigma^2\}$$

as the union of all $(\mathbf{x}, y)$ belonging to one of the $w_{\boldsymbol{\beta},\sigma^2}$-stripes for $(\boldsymbol{\beta}, \sigma^2) \in C$. Consistency of $f_{\mathbf{Z}_n}$ for $f_P$ within $C$ requires the following assumptions:

$$\forall (\boldsymbol{\beta}, \sigma^2) \in C : P\{(y - \mathbf{x}'\boldsymbol{\beta})^2 = c\sigma^2\} = 0, \tag{9}$$

$$\forall \mathbf{z} \in \mathbb{R}^{p+1} \backslash \{0\} : P(\{\mathbf{x}'\mathbf{z} = 0\} \cap V(C)) = 0, \tag{10}$$

$$\begin{aligned} &Py^2 1[(\mathbf{x}, y) \in V(C)] < \infty, \\ &P||\mathbf{x}||^2 1[(\mathbf{x}, y) \in V(C)] < \infty, \end{aligned} \tag{11}$$

$$\inf_{(\boldsymbol{\beta},\sigma^2) \in C} Pw_{\boldsymbol{\beta},\sigma^2} > 0. \tag{12}$$

Assumptions (9) and (10) are fulfilled if $P$ is Lebesgue-dominated. Finiteness of $Py^2$ and $P||\mathbf{x}||^2$ suffices for (11). Eq. (9) and the moment conditions (11) are needed to ensure the continuity of $f_P$. Eq. (10) prevents the covariate matrix from getting collinear inside of $V(C)$. Assumption (12) together with (9) forces $C$ to be bounded away from $\sigma^2 = 0$. The latter suffices for (12) to hold if $P$ has a non-vanishing Lebesgue-density, since $C$ is compact. Eq. (12) is necessary since FPC analysis deals with arbitrary small subsets of the data, and increasing $n$ does not prevent the occurrence of very small data subsets such that their local estimators of regression and error variance lie far from their theoretical values.

**Lemma 5.1.** *If* (9)–(12) *hold for a compact* $C \subset \mathbb{R}^{p+1} \times \mathbb{R}_0^+$, *then for all* $\kappa > 0$

$$P^\infty\{\exists n_0\ \forall n > n_0, (\boldsymbol{\beta}, \sigma^2) \in C : ||f_{\mathbf{Z}_n}(\boldsymbol{\beta}, \sigma^2) - f_P(\boldsymbol{\beta}, \sigma^2)|| < \kappa\} = 1.$$

Proofs are given in the Appendix.

This means that for such a $C$, which can be arbitrary large as long as it is compact and bounded away from $\sigma^2 = 0$, LS-FPCVs may occur only outside of $C$ or where $f_P(\boldsymbol{\beta}, \sigma^2)$ is close to $(\boldsymbol{\beta}, \sigma^2)$ for large enough $n$:

**Corollary 1.** *Let* $\kappa > 0$. *Let* $C$ *fulfill the assumptions of Lemma* 5.1. *Then for large enough* $n$, $P^\infty$*-a.s., no LS-FPCV* $\mathbf{w}_{\mathbf{Z}_n, \boldsymbol{\beta}, \sigma^2}$ *exists with* $(\boldsymbol{\beta}, \sigma^2) \in C$ *and* $||f_P(\boldsymbol{\beta}, \sigma^2) - (\boldsymbol{\beta}, \sigma^2)|| \geqslant \kappa$.

The corollary follows directly from Lemma 5.1.

A further assumption is required to show the existence of consistent sequences of LS-FPCVs for LS-FPCIs: Suppose

$$\exists \text{ LS-FPCI } w_{\boldsymbol{\beta}_0, \sigma_0^2} \text{ w.r.t. } P,\ \sigma_0^2 > 0. \tag{13}$$

(If there exists such LS-FPCI with $\sigma_0^2 = 0$, then $Pw_{\boldsymbol{\beta}_0, 0} > 0$. For large enough $n$ there are enough points $(\mathbf{x}, y)$ with $(y - \mathbf{x}'\boldsymbol{\beta}_0)^2 = 0$, $P^\infty$-a.s., so that $\mathbf{w}_{\mathbf{Z}_n, \boldsymbol{\beta}_0, 0}$ is an LS-FPCV. That is, in this case a consistent sequence of LS-FPCVs exists.)

It will be assumed that $\exists \varepsilon_0 > 0$, $1 > \alpha \geqslant 0$:

$$\forall 0 < \varepsilon \leqslant \varepsilon_0 : (\boldsymbol{\beta}, \sigma^2) \in B_\varepsilon(\boldsymbol{\beta}_0, \sigma_0^2) \Rightarrow f_P(\boldsymbol{\beta}, \sigma^2) \in B_{\alpha\varepsilon}(\boldsymbol{\beta}_0, \sigma_0^2). \tag{14}$$

This assumption is needed to force $f_{Z_n}(\boldsymbol{\beta}, \sigma^2)$, where $(\boldsymbol{\beta}, \sigma^2)$ is close to $(\boldsymbol{\beta}_0, \sigma_0^2)$, into a shrinking neighborhood of $(\boldsymbol{\beta}_0, \sigma_0^2)$. Let $C := B_{\varepsilon_0}(\boldsymbol{\beta}_0, \sigma_0^2)$. Eq. (14) follows immediately if

$$f_P(C) \subseteq C,\ \ 1 > \alpha \geqslant 0 : \forall (\boldsymbol{\beta}_1, \sigma_1^2), (\boldsymbol{\beta}_2, \sigma_2^2) \in C :$$
$$||f_P(\boldsymbol{\beta}_1, \sigma_1^2) - f_P(\boldsymbol{\beta}_2, \sigma_2^2)|| \leqslant \alpha ||(\boldsymbol{\beta}_1, \sigma_1^2) - (\boldsymbol{\beta}_2, \sigma_2^2)||, \tag{15}$$

i.e., contractivity of $f_P$ within $C$ as needed for Banach's Fixed Point Theorem that guarantees the existence of a fixed point within $C$ (but only for $f_P$, not for the non-continuous $f_{\mathbf{Z}_n}$). See Section 6 for a discussion of cases where this is fulfilled.

**Theorem 5.1.** *Assume* (13), (14) *and* (9)–(11) *for* $C = B_{\varepsilon_0}(\boldsymbol{\beta}_0, \sigma_0^2)$. *Then*,

$$P^\infty\{\forall n > p+1 \exists \mathbf{w}_{\mathbf{Z}_n, \boldsymbol{\beta}_n, \sigma_n^2} \text{ LS-FPCV w.r.t. } \mathbf{Z}_n:$$
$$\lim_{n \to \infty} (\boldsymbol{\beta}_n, \sigma_n^2) = (\boldsymbol{\beta}_0, \sigma_0^2)\} = 1$$

**Remark 2.** Conditions (14) and (15) are equivariant under data transformations of the form $D(\mathbf{x}, y) = (\mathbf{A}\mathbf{x}, ay + \mathbf{x}'b)$ insofar as they hold for the distribution $P^D$ of the transformed data and the corresponding fixed points of $f_{P^D}$ with respect

to the norm

$$||\mathbf{z}||_{(\mathbf{B}^{-1})'\mathbf{B}^{-1}} := \mathbf{z}'(\mathbf{B}^{-1})'\mathbf{B}^{-1}\mathbf{z},$$

where $\mathbf{B} := \begin{pmatrix} a(\mathbf{A}^{-1})' & 0 \\ 0 & a^2 \end{pmatrix}$ is assumed to be invertible.

## 6. LS-fixed point cluster indicators of some contamination and mixture models

This section gives some results concerning distributions of type (7). First, the existence and uniqueness of an LS-FPCI in the case $\varepsilon = 0$ is shown. Corollary 2 and Lemma 6.1 (giving conditions for an LS-FPCI to fulfill the assumptions of Theorem 5.1) allow $\varepsilon > 0$, but require $P^*$ to give mass 0 to some neighborhood of $\{y = \mathbf{x}'\boldsymbol{\beta}_1\}$. This does not hold for mixtures of more than one regression with normal distributed errors. An example of a normal mixture ($p = 0$) is discussed at the end of the section.

In the case $\varepsilon = 0$, $P = P_{\boldsymbol{\beta}_1,\sigma_1^2,G}$ is a homogeneous linear regression distribution. Consequently there is only one LS-FPCI. Its parameters are $\boldsymbol{\beta}_1$ and $k\sigma_1^2$, where $k\sigma_1^2$ is the variance of the truncated normal distribution belonging to the LS-FPCI. For example, $c = 10$ yields $k = 0.9815$, $c = 6.635$ yields $k = 0.9001$.

**Theorem 6.1.** *Let $c > 3$. $w_{\boldsymbol{\beta}_1,k\sigma_1^2}$ is the unique LS-FPCI w.r.t. $P = P_{\boldsymbol{\beta}_1,\sigma_1^2,G}$, where $k$ is the unique zero of*

$$h(k) := 1 - k - \frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck}) - \Phi(-\sqrt{ck})}.$$

The theorem leads easily to the existence of a suitable LS-FPCI in the contamination model with $\varepsilon > 0$, if there is no overlap between the LS-FPCI of the component $P_{\boldsymbol{\beta}_1,\sigma_1^2,G}$ and $P^*$:

**Corollary 2.** *$w_{\boldsymbol{\beta}_1,k\sigma_1^2}$ is LS-FPCI w.r.t. $P$ defined by* (7) *with $P_0 = P_{\boldsymbol{\beta}_1,\sigma_1^2,G}$, if*

$$P^* w_{\boldsymbol{\beta}_1,k\sigma_1^2} = 0. \tag{16}$$

**Proof.** $P_{w_{\boldsymbol{\beta}_1,k\sigma_1^2}}$ does not change between Theorem 6.1 and Corollary 2. □

The uniqueness of the LS-FPCI is lost in this case. This is reasonable since $P^*$ may generate clusters elsewhere. Eq. (16) means that $P^*$ has to generate outliers w.r.t. $P_{w_{\boldsymbol{\beta}_1,k\sigma_1^2}}$ with probability 1.

Now, conditions will be given to ensure that the LS-FPCI of Corollary 2 fulfills the assumptions of the consistency theorem.

**Lemma 6.1.** *Let* $c>3$, $P=(1-\varepsilon)P_{\boldsymbol{\beta}_1,\sigma_1^2,G}+\varepsilon P^*$, $1>\varepsilon\geqslant 0$, *where*

$$G||\mathbf{x}||^3<\infty,\ \forall \mathbf{a}\neq 0: G\{\mathbf{a}'\mathbf{x}=0\}=0, \tag{17}$$

$$\exists \varepsilon_1>0: P^*(V(B_{\varepsilon_1}(\boldsymbol{\beta}_1,k\sigma_1^2)))=0, \tag{18}$$

*where* $k>0$ *is defined as in Theorem* 6.1 *and fulfills furthermore*

$$k>1-\frac{2}{c-1}. \tag{19}$$

*Then the assumptions of Theorem* 5.1 *are fulfilled with* $\boldsymbol{\beta}_0=\boldsymbol{\beta}_1$, $\sigma_0^2=k\sigma_1^2$.

Eq. (19) can be verified numerically for given $c$ and holds for all values which are applied in this paper.

In the case $p=0$, the function $f_P$ can be evaluated and visualized numerically for normal mixtures. Fig. 3 gives an example for the LS-FPCIs of a normal mixture, namely $P=\frac{1}{2}\mathcal{N}_{0,1}+\frac{1}{2}\mathcal{N}_{5,0.25}$ with $c=6.635$. There are five LS-FPCIs. By visual inspection of the function $f_P$ it can be seen that three of them fulfill (15), They are shown as fat lines in Fig. 3. Two of them correspond to the two mixture components. The third one corresponds to the bulk of the mass of the whole distribution and could only be considered as "homogeneous" compared to any added gross outliers. Such an LS-FPCI exists almost always and illustrates that the definition does not enforce FPCs to have a Gaussian shape. They are only homogeneous *compared* to what is far away in the dataset.

It can be shown that the fixed points of $f_P$ leading to the two non-fat intervals not only violate (15), but are in fact "repulsive". This means that a fixed point algorithm applied to $f_P$ never will converge to these fixed points unless they are used as starting

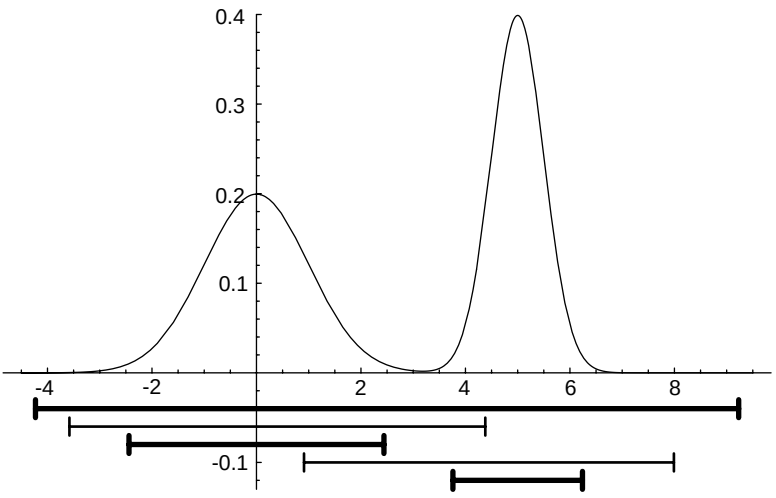


Fig. 3. $P=\frac{1}{2}\mathcal{N}_{0,1}+\frac{1}{2}\mathcal{N}_{5,0.25}$: p.d.f. with FPCs (fat intervals: (15) fulfilled).

values. The corresponding LS-FPCIs, which cannot be interpreted as proper "clusters", may therefore not be expected to lead to FPCVs in datasets when the fixed point algorithm is applied to find the FPCs. Further examples are given in [18,20].

## 7. Implementation of the procedure

An exhaustive search for all LS-FPCVs of a given dataset is impossible unless $n$ is very small. In this section an implementation of a procedure is described to find all "substantial" FPCVs with high probability. The implementation and the choice of the needed constants is discussed in detail in [30].

The basic procedure is simple:

1. Choose the number of algorithm runs $i_{n,p}$ and the tuning constant $c$.
2. Repeat $i_{n,p}$ times: Generate a subset indicator $\mathbf{w}^0$ with $n(\mathbf{w}^0) = p + 2$ randomly and apply the FPA until convergence.

Applying the basic procedure, one may observe that the number of found FPCs is often larger than one would like to have for the clarity of the interpretation, unless $n$ is very large or $i_{n,p}$ is so small that the result of the analysis depends strongly upon chance.

The following implementation aims to exclude FPCs which are too small, too unstable or too similar to other FPCs. For a justification and simulations, which indicate a reasonable performance of the whole procedure, see [30].

*Step* 1: Choose $c$, say $c = 6.635$ (see Section 4; simulations indicate that $c$ should be larger for small $n$ and large $p$), $i_{\min} = 3$, $s_{\text{cut}} = 0.85$, $n(\mathbf{w}^0) = p + 2$, $n_{\min} = \frac{n}{5}$. $n_{\min}$ is the minimum size of an FPCV for which an appropriate starting constellation occurs $i_{\min}$ times with a probability of $\geqslant 0.95$. $s_{\text{cut}}$ is a minimum value of a similarity between FPCs so that they are considered as "corresponding to the same structure". The similarity measure between the indicator vectors $\mathbf{v}$ and $\mathbf{w}$ of subsets of a dataset is defined as follows:

$$s_*(\mathbf{v}, \mathbf{w}) := \frac{2|\{i : v_i w_i = 1\}|}{|\{i : v_i = 1\}| + |\{i : w_i = 1\}|}. \tag{20}$$

*Step* 2: Compute $i_{n,p}$ according to

$$i_{n,p} := \min\left\{ i : \mathrm{QB}\left( i, \frac{\binom{n_{\min}}{n(\mathbf{w}^0)}}{\binom{n}{n(\mathbf{w}^0)}}; 0.05 \right) < i_{\min} \right\}, \tag{21}$$

where $\mathrm{QB}(n, p; \alpha)$ denotes the $\alpha$-quantile of the Binomial $(n, p)$-distribution. The idea here is that for a given LS-FPCV of size of at least $n_{\min}$ the probability should be

larger than 0.95 that all points of the starting configuration come from it at least $i_{\min}$ times.

*Step* 3: Repeat $i_{n,p}$ times: Generate a subset indicator $\mathbf{w}^0$ by random and apply the FPA. Store all found FPCVs $\mathbf{w}$ with $n(\mathbf{w}) \geqslant n_{\min}$.

*Step* 4: Compute the similarities for each pair of FPCVs according to (20).

*Step* 5: Compute the Single Linkage clusters of index $s_{\text{cut}}$ of the FPCVs, i.e., the connectivity components of the graph of FPCVs where two FPCVs are linked if their similarity exceeds $s_{\text{cut}}$. An algorithm is given by Cormen et al. [7, p. 477].

*Step* 6: For every cluster of FPCVs, call the FPCV which is found most often the "representative" FPCV. Discard all clusters of FPCVs whose members were found fewer than $i_{\min}$ times.

Unfortunately, $i_{n,p}$ increases exponentially with $p$ if chosen according to (21). Thus, $p \geqslant 4$ results in very large computation times.

## 8. Application to the Old Faithful data

Data on the duration of eruptions and the waiting time between the eruptions of the Old Faithful Geyser in the Yellowstone National Park have been discussed in several publications on the basis of data from various time periods. A literature overview, as well as the dataset analyzed here, can be found in [2]. These data were collected in August 1985. Measurements are in minutes. They are shown in Fig. 1.

The duration of an eruption of the geyser is modeled here as dependent upon the waiting time since the previous eruption. There seem to be at least two different groups of dependency, corresponding to the eruptions with lower and higher duration. The latter group shows a moderately decreasing tendency for increasing waiting times.

Some authors (e.g. [6]) model the duration of an eruption as an independent covariate for the subsequent waiting time. Their approach does not reveal any differences between groups. There are no publications up to now that address clustering and dependency between successive events at the same time. Azzalini and Bowman [2] analyze the data with time series models. They *assume* for their analysis that there are two different patterns of dependency, while I use the dataset to illustrate how to *find* such kind of heterogeneity.

The data show some other features: There is a clear outlier with a duration value smaller than 1. The probability for a long eruption was clearly larger if the waiting time had been short, i.e., the assignment to the two groups is dependent upon the independent variable. There are 53 points with duration$= 4$ exactly, and there are about 20 points with duration$= 2$. This is due to inexact observations during the night, which were coded as 2 (short eruption), 3 (medium length eruption, only once) and 4 (long eruption) by Azzalini and Bowman.

FPC analysis was applied according to the procedure described in Section 7, i.e., $c = 6.635$, $i_{n,p} = 809$. This resulted in eight FPCs. There were six Single Linkage

groups of FPCs, and four of them were found three times or more. I concentrate on the interpretation of the four representative FPCs.

The whole dataset was found 521 times as an FPC. It has been discussed previously (Section 4, Fig. 3) that there is usually an FPC corresponding to (almost) the whole dataset. This FPC can be expected to be found often, namely always if the points of the starting configuration do not all belong to the same smaller cluster. This is an artifact of the method and has to be taken into account in order to interpret the results.

The other representative FPCs are more interesting. The Single Linkage group of the second one was found 217 times. It consists of the circles together with the crosses of Fig. 2 of Section 3 and corresponds to the group with the longer durations. It excludes the points with the two largest durations as outliers as well as most of the points with medium duration of the eruption. The Single Linkage group of the third representative FPC was found 31 times. It contains the points denoted by triangles in Fig. 2 and corresponds to the group with the shorter durations, excluding the outlier with duration smaller than 1. The points with duration$=4$, denoted by crosses, form the fourth representative FPC, which was found 8 times.

The second, third and fourth representative FPCs give a good description of the main features of the dataset. The possibility of overlapping FPCs is useful here, since an interpretation of the points with duration $=4$ as its own cluster is reasonable (''group of inexactly observed long eruptions'') *as well as* an interpretation of them as a part of the larger ''long duration''-group. The points with duration $=2$ form an FPC as well, but it was not found often enough during the iterations, since its number of points is too small.

I applied two mixture model methods to the dataset as well. The first method is the maximum likelihood clusterwise linear regression mixture (MLCLR) estimator computed by the EM-algorithm as explained by DeSarbo and Cron [12] with estimation of the number of components by use of the Consistent Akaike Information Criterion as recommended by Wedel and DeSarbo [29]. I implemented it in the freeware `R`. The result (with classification of the observations according to maximum a posteriori membership probability) is shown on the left side of Fig. 4. The solution suffers from an implicit assumption of the linear regression mixture model, namely that the proportion of the mixture components has to be the same for all values of the independent variable [19]. For example, the cluster consisting of circles corresponds roughly to the eruptions with lower durations, but contains also some points with low waiting times, which from graphical inspection should be assigned to another cluster.

The other method is the software `mclust` for model-based Gaussian clustering with noise (MBGCN) described by Fraley and Raftery [14], which was used for the detection of linearly shaped clusters in the presence of noise by DasGupta and Raftery [9]. The method performs maximum likelihood estimation in a mixture of multivariate normal distributions. Various models are defined by different constraints on the covariance matrices, and an optimal model and an optimal number of clusters are chosen by use of the Bayesian information criterion. The method allows for noise by the introduction of a Poisson process mixture

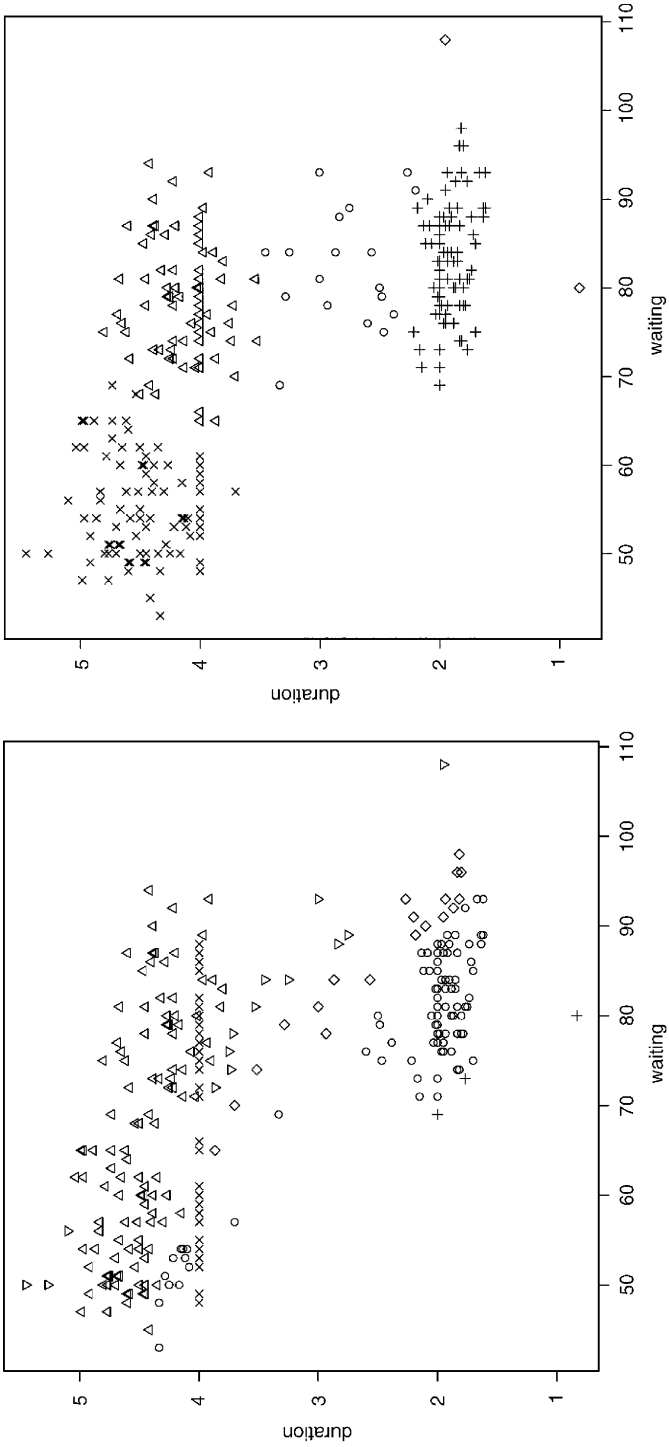


Fig. 4. Old Faithful Geyser data with partition of MLCLR (left) and MBGCN (right, diamonds denoting estimated noise).

component. An initial noise estimate is needed. For this purpose I used the software `NNclean` by Byers and Raftery [3] as discussed by Fraley and Raftery [14]. Both programs were designed for `Splus` and ported to `R`. The result can be seen on the right side of Fig. 4. `mclust` chooses four clusters plus a noise component of two points (diamonds). The points with large and small duration are properly separated. The points between the high and low duration clusters were estimated as a single mixture component, but they do not look like a sample from a multivariate normal. The points with long duration are divided into parts with larger and smaller waiting times. This seems to be reasonable by graphical inspection. However, Azzalini and Bowman [2] give geological evidence for the existence of two distinct patterns of eruptions, corresponding to the second and third representative FPC, while there are no arguments for breaking the "long duration"-group into two parts as in the MBGCN-solution. Generally, it is questionable to analyze data with mixture models of unstructured distributions if there is background knowledge that explains a possible clustering by different relations between distinguished independent and dependent variables.

The FPC solution seems to be most useful for this data. However, it has to be noted that the algorithms for the mixture maximum likelihood methods are also sensitive to some parameter choices, namely the choice of starting configurations, a convergence criterion, and a lower bound for the error variance, covariance determinant, respectively. The latter is necessary, because otherwise the likelihood would be unbounded [24, Section 3.8]. The ability of the method to find the exact lines with duration 2 and 4 as clusters depends crucially on the choice of this lower bound. My choice of $10^{-6}$ for MLCLR was small enough that the duration-4-line was found. I think that the duration-2-line with fewer points would have needed another starting configuration. MBGCN computed with the unchanged parameter values of `mclust` did not succeed to find these lines. There may be even better parameter choices than mine.

## 9. Conclusion

A new concept to define clusters was presented in this paper: An FPC is a data subset that does not contain any outlier, but w.r.t. which all other points are outliers. The basic idea for FPC analysis is to compute iteratively reweighted estimators for which all outliers have zero weight, as for redescending M-estimators. FPC analysis was developed here for clusterwise linear regression, but it may be adapted to other clustering problems. FPCs are not necessarily exhaustive, they may intersect and include each other and they are locally defined, i.e., the FPC property of a data subset does not depend upon distant parts of the dataset.

The existence and consistent estimability of theoretical LS-FPCs of certain probability models of interest was investigated. FPC analysis is not meant as a procedure which is optimal with respect to particular reference models and target functions. It should be a data analytic tool which may be valuable under various

deviations from the standard models of the model based CA. However, I tried to give it a solid stochastic foundation. The consistency results have their limitations. In particular, it is sometimes not easy to verify the assumptions for $p>0$. Note, however, that to my knowledge at the moment all consistency proofs for the estimation of the parameters of a linear regression mixture suffer from not taking the question of identifiability adequately into account (see [19]). The FPC theory has no problems to deal with the situations of non-identifiability of the regression parameters discussed in [19], because FPC analysis does not fit a global model, and consequently it needs not to decide between distinct parameterizations leading to the same global distribution.

The Old Faithful Geyser dataset turned out as an example where FPC analysis can be applied successfully, while ordinary mixture model methods have problems. Three features of FPC analysis are useful here:

1. Outliers do not need to fit in any parametric model.
2. FPCs may intersect.
3. FPC analysis treats data subsets with zero error variance in a natural way.

I compared FPC analysis with MLCLR and MBGCN by simulations [30]. The result is that the mixture methods perform better if their model assumptions are fulfilled, but FPC analysis can outperform MLCLR in datasets, where the cluster assignment depends upon the independent variables, and it can be better than MBGCN under non-normal distributions of the independent variables. The simulations treat the recovery of mixture components, but it has to be noted that FPC analysis is essentially not a method to find mixture components, but rests on its own cluster definition.

A drawback of FPC analysis in its present implementation is the large computing time for higher dimensions. However, the FPC algorithm of Section 3.2 can be applied easily in any dimension, if subject-matter knowledge or graphical inspection yield candidate data subsets, from which the algorithm can be started, see [17].

An R-module and a C-software for LS-FPC analysis, and the reference Hennig [20] can be obtained from `http://www.math.uni-hamburg.de/home/hennig/`

## Appendix A. Proofs

### *A.1. Some useful results for the following proofs*

**Assumption.** Let $Q$ be a measure on $(\mathbb{R}^{p+2}, \mathbb{B}^{p+2})$, $M \subseteq \mathbb{R}^{p+1} \times \mathbb{R}^+$, where

$$Qy^2 < \infty, \quad Q||\mathbf{x}||^2 < \infty, \quad (Q\mathbf{x}\mathbf{x}')^{-1} \text{ exists}, \tag{A.1}$$

$$Q\{(y - \mathbf{x}'\boldsymbol{\beta})^2 = c\sigma^2\} = 0 \ \forall (\boldsymbol{\beta}, \sigma^2) \in M, \tag{A.2}$$

$$(Q\mathbf{x}\mathbf{x}'\mathbf{w}_{\boldsymbol{\beta},\sigma^2}(\mathbf{x}, y))^{-1} \text{ exists } \forall (\boldsymbol{\beta}, \sigma^2) \in M. \tag{A.3}$$

**Proposition A.1.** $\arg\min_{\beta} Q(y-\mathbf{x}'\boldsymbol{\beta})^2 = (Q\mathbf{x}\mathbf{x}')^{-1}Q\mathbf{x}y$ *exists uniquely under* (A.1).

**Proposition A.2.** *Let* $l_1(a,\mathbf{b}) := Qv(\mathbf{x},y)1[(y-\mathbf{x}'\mathbf{b})^2 \leqslant ca^2]$ *where* $v:\mathbb{R}^{p+2}\mapsto\mathbb{R}^q$, $Q||v(\mathbf{x},y)|| < \infty$. $l_1$ *is continuous on* $M$ *under* (A.2).

$l_2(a_1,\mathbf{b}_1,a_2,\mathbf{b}_2) := Q1[(\mathbf{x}'\mathbf{b}_2)^2 > a_2^2]1[(y-\mathbf{x}'\mathbf{b}_1)^2 \leqslant ca_1^2]$ *is continuous in* $(a_1,\mathbf{b}_1,a_2,\mathbf{b}_2)\in M\times\mathbb{R}^{p+2}$ *under* (A.2) *and* (A.3).

$f_Q$ *is continuous on* $M$ *under* (A.1)–(A.3).

The proofs are straightforward.

### *A.2. Further preparations for the proof of Lemma 5.1*

Here are some useful results proven e.g. in [28]:

Let $\mathscr{F}$ be a class of measurable functions $\mathbb{R}^k\mapsto\mathbb{R}$. For a real-valued function $h$, $|h|$ denotes the supremum norm. A measurable function $F$ with $|f|\leqslant F\ \forall f\in\mathscr{F}$ is called *Q-finite envelope* if $Q$ is a measure with $QF<\infty$.

**Definition A.1.** For $\varepsilon>0$ and a measure $Q$ on $(\mathbb{R}^d,\mathbb{B}^d)$, the *covering number* $N(\varepsilon,\mathscr{F},Q)$ is the minimum number of balls $\{g: Q|g-f|<\varepsilon\}$ needed to cover $\mathscr{F}$. The centers $f$ need not belong to $\mathscr{F}$.

**Definition A.2.** $\mathscr{F}$ is called *permissible* if it can be indexed by some $T$ with Borel-$\sigma$-field $\mathscr{B}$ in such a way that $f(\bullet,\bullet)$ is $\mathbb{B}^k\otimes\mathscr{B}$-measurable and $T$ is an analytic subset of some compact metric space.

**Theorem A.1.** *Let* $\mathscr{F}$ *be permissible with P-finite envelope* $F$. *If*

$$\forall\varepsilon>0:\ \log N(\varepsilon,\mathscr{F},P_n)=o_P(n), \tag{A.4}$$

*then*

$$\sup_{f\in\mathscr{F}}|P_nf-Pf|\to 0\ \ P^\infty\text{-a.s.}$$

**Definition A.3.** Let $\mathscr{C}$ be a collection of subsets of a set $S$. $\mathscr{C}$ is said to *shatter* a set $S_*\subset S$ if every subset of $S_*$ can be formed as $S_*\cap C,\ C\in\mathscr{C}$. $\mathscr{C}$ is called a *Vapnik–Chervonenkis* (*VC*)*-class* with index $V(\mathscr{C})\in\mathbb{N}$, if $\mathscr{C}$ shatters no subset of $S$ with $V(\mathscr{C})$ elements.

**Proposition A.3.** $\mathscr{F}$ *fulfills* (A.4), *if the set of subgraphs* $\{(\mathbf{x},t)\in\mathbb{R}^{k+1}: t<f(\mathbf{x})\}$, $f\in\mathscr{F}$, *is a VC-class. Such* $\mathscr{F}$ *is itself called VC-class.*

Below the VC-class property is claimed for some sets. All such assertions made here are easily proven by help of Chapter 2 of van der Vaart and Wellner [28]. For details see [20].

Further propositions are needed for the proof of Lemma 5.1. For $(\mathbf{t},\boldsymbol{\beta},\sigma^2)\in\mathbb{R}^{p+1}\times\mathbb{R}^{p+1}\times\mathbb{R}_0^+$ define

$$f_{\mathbf{t},\boldsymbol{\beta},\sigma^2}(\mathbf{x},y) := (y-\mathbf{x}'\mathbf{t})^2 1[(y-\mathbf{x}'\boldsymbol{\beta})^2\leqslant c\sigma^2],\quad (\mathbf{x},y)\in\mathbb{R}^p\times\{1\}\times\mathbb{R}.$$

**Proposition A.4.** *Under the assumptions of Lemma 5.1, define* $\sigma_*^2 := \max\{\sigma^2:(\boldsymbol{\beta},\sigma^2)\in C\}$. $\forall\eta>0\ \exists d_{C,\eta}<\infty$ *such that* $||\mathbf{t}||>d_{C,\eta}\Rightarrow\inf_{(\boldsymbol{\beta},s^2)\in C}Pf_{\mathbf{t},\boldsymbol{\beta},s^2}>c\sigma_*^2+\eta$ *and,* $P^\infty$*-a.s. for large enough* $n$, $\inf_{(\boldsymbol{\beta},s^2)\in C}P_nf_{\mathbf{t},\boldsymbol{\beta},s^2}>c\sigma_*^2+\eta$.

**Proof.** $S_p := \{\mathbf{x}\in\mathbb{R}^{p+1}:||\mathbf{x}||=1\}$. Show that there exists $\tau>0$ such that

$$a_P := \inf_{(\boldsymbol{\beta},\sigma^2)\in C,\mathbf{z}\in S_p} P(L_{\boldsymbol{\beta},\sigma^2,\tau,\mathbf{z}})>0,$$
$$L_{\boldsymbol{\beta},\sigma^2,\tau,\mathbf{z}} := \{|\mathbf{x}'\mathbf{z}|>\tau\}\cap\{w_{\boldsymbol{\beta},\sigma^2}=1\}. \tag{A.5}$$

Suppose that (A.5) does not hold. Then, because of the compactness of $S_p$ and $C$, there is a sequence $(\boldsymbol{\beta}_n,\sigma_n^2,\tau_n,\mathbf{z}_n)_{n\in\mathbb{N}}$ with

$$(\boldsymbol{\beta}_n,\sigma_n^2,\tau_n,\mathbf{z}_n)\to(\boldsymbol{\beta}_0,\sigma_0^2,0,\mathbf{z}_0),(\boldsymbol{\beta}_0,\sigma_0^2)\in C,\quad \mathbf{z}_0\in S_p,$$

and $P(L_{\boldsymbol{\beta}_n,\sigma_n^2,\tau_n,\mathbf{z}_n})\to 0$. Use Proposition A.2 to get $P(L_{\boldsymbol{\beta}_0,\sigma_0^2,0,\mathbf{z}_0})=0$. But this contradicts (12) because of $P(\{|\mathbf{x}'\mathbf{z}_0|=0\}\cap\{w_{\boldsymbol{\beta}_0,\sigma_0^2}=1\})=0$ by (10). Therefore (A.5).

Further get

$$a_{P_n} = \inf_{(\boldsymbol{\beta},\sigma^2)\in C,\ \mathbf{z}\in S_p} P_n(L_{\boldsymbol{\beta},\sigma^2,\tau,\mathbf{z}})\to a_P>0\quad P^\infty\text{-a.s.}$$

since the sets $\{|\mathbf{x}'\mathbf{z}|>\tau\}=\{\mathbf{x}'\mathbf{z}\mathbf{z}'\mathbf{x}-\tau^2>0\}$, $\mathbf{z}\in S_p$ form a VC-class, and so do the intersections with the sets $\{w_{\boldsymbol{\beta},\sigma^2}=1\}$, $(\boldsymbol{\beta},\sigma^2)\in C$ and their indicator functions, to which Theorem A.1 can be applied; permissibility is obvious. For $\mathbf{t}\in\mathbb{R}^{p+1}$, $(\boldsymbol{\beta},\sigma^2)\in C$, $(\mathbf{x},y)\in L_{\boldsymbol{\beta},\sigma^2,\tau,\frac{\mathbf{t}}{||\mathbf{t}||}}$, get $|\mathbf{x}'\mathbf{t}|=|\mathbf{x}'\frac{\mathbf{t}}{||\mathbf{t}||}|||\mathbf{t}||\geqslant||\mathbf{t}||\tau$ and hence for $Q=P$ and $P^\infty$-a.s. for sufficiently large $n$ for $Q=P_n$:

$$\begin{aligned}
Qf_{\mathbf{t},\boldsymbol{\beta},\sigma^2} &= \int (y-\mathbf{x}'\mathbf{t})^2 1[(y-\mathbf{x}'\boldsymbol{\beta})^2\leqslant c\sigma^2]\,dQ\\
&\geqslant \int (y-\mathbf{x}'\mathbf{t})^2 1[L_{\boldsymbol{\beta},\sigma^2,\tau,\frac{\mathbf{t}}{||\mathbf{t}||}}]\,dQ\\
&\geqslant \int |\mathbf{t}'\mathbf{x}|(|\mathbf{t}'\mathbf{x}|-2|y|)1[L_{\boldsymbol{\beta},\sigma^2,\tau,\frac{\mathbf{t}}{||\mathbf{t}||}}]\,dQ\\
&\geqslant ||\mathbf{t}||\tau(||\mathbf{t}||\tau a_Q-2Q|y|),
\end{aligned}$$

which exceeds $c\sigma_*^2$ for sufficiently large $||\mathbf{t}||$ since $a_Q>0$ and $Q|y|<\infty$. Existence of $d_{C,\eta}$ follows.

**Corollary A.1.** *Let* $\eta>0$, $(\boldsymbol{\beta},\sigma^2)\in C$. *For* $Q=P$ *and* $P^\infty$*-a.s. for* $Q=P_n$, $n$ *large enough*

$$\inf_{||\mathbf{t}||\geqslant d_{C,\eta}} Qf_{\mathbf{t},\boldsymbol{\beta},\sigma^2} > \arg\min_{\mathbf{t}} Qf_{\mathbf{t},\boldsymbol{\beta},\sigma^2} + \eta.$$

**Proof.** $Qf_{\boldsymbol{\beta},\boldsymbol{\beta},\sigma^2}\leqslant c\sigma_*^2$ by definition of $f_{\mathbf{t},\boldsymbol{\beta},\sigma^2}$.

**Proposition A.5.** *Under the assumptions of Lemma* 5.1, $\forall\kappa>0$:

$$\inf_{(\boldsymbol{\beta},\sigma^2)\in C}\ \inf_{||\mathbf{t}-\beta_P(\boldsymbol{\beta},\sigma^2)||\geqslant\kappa} (Pf_{\mathbf{t},\boldsymbol{\beta},\sigma^2} - Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2})>0.$$

**Proof.** Suppose that the proposition does not hold, i.e., there is some sequence $(\mathbf{t}_m,\boldsymbol{\beta}_m,\sigma_m^2)_{m\in\mathbb{N}}$ where $||\mathbf{t}_m-\beta_P(\boldsymbol{\beta}_m,\sigma_m^2)||>\kappa>0$ and

$$|Pf_{\mathbf{t}_m,\boldsymbol{\beta}_m,\sigma_m^2} - Pf_{\beta_P(\boldsymbol{\beta}_m,\sigma_m^2),\boldsymbol{\beta}_m,\sigma_m^2}|\to 0.$$

$(\mathbf{t}_m,\boldsymbol{\beta}_m,\sigma_m^2)_{m\in\mathbb{N}}$ has a compact domain since $(\boldsymbol{\beta}_m,\sigma_m^2)\in C$ and $||\mathbf{t}_m||\leqslant d_{C,\kappa}$ by Corollary A.1 for large $m$. Hence the sequence can be chosen convergent to some $(\mathbf{t}_0,\boldsymbol{\beta}_0,\sigma_0^2)\in\{||\mathbf{t}||\leqslant d_{C,\eta}\}\times C$ where $|\mathbf{t}_0-\beta_P(\boldsymbol{\beta}_0,\sigma_0^2)|\geqslant\kappa$. $Pf_{\mathbf{t},\boldsymbol{\beta},\sigma^2}$ is continuous in $(\mathbf{t},\boldsymbol{\beta},\sigma^2)$ under (9) and (11) by Proposition A.2, thus

$$Pf_{\mathbf{t}_m,\boldsymbol{\beta}_m,\sigma_m^2}\to Pf_{\mathbf{t}_0,\boldsymbol{\beta}_0,\sigma_0^2} = Pf_{\beta_P(\boldsymbol{\beta}_0,\sigma_0^2),\boldsymbol{\beta}_0,\sigma_0^2}.$$

In contradiction to that, $\arg\min_{\mathbf{t}} Pf_{\mathbf{t},\boldsymbol{\beta}_0,\sigma_0^2}$ is uniquely defined because of (10) and (11) (Proposition A.1). This proves the proposition. □

**Proof of Lemma 5.1.** Define for $\eta>0$

$$\mathscr{F}_C := \{f_{\mathbf{t},\boldsymbol{\beta},\sigma^2} : ||\mathbf{t}||\leqslant d_{C,\eta},\ (\boldsymbol{\beta},\sigma^2)\in C\}.$$

$\mathscr{F}_C$ is permissible by its parameterization. $\mathscr{F}_C$ has $P$-finite envelope $F_C(\mathbf{x},y) := y^2+2|y|||\mathbf{x}||d_{C,\eta}+||\mathbf{x}||^2d_{C,\eta}^2$ and is a VC-class. Theorem A.1 yields, $P^\infty$-a.s.,

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C,||\mathbf{t}||\leqslant d_{C,\eta}} |P_nf_{\mathbf{t},\boldsymbol{\beta},\sigma^2} - Pf_{\mathbf{t},\boldsymbol{\beta},\sigma^2}|\to 0. \tag{A.6}$$

By definition $\beta_{\mathbf{Z}_n}(w_{\boldsymbol{\beta},\sigma^2}) = \arg\min_{\mathbf{t}} P_nf_{\mathbf{t},\boldsymbol{\beta},\sigma^2}$. For sufficiently large $n$, the arg min can be taken over $\{||\mathbf{t}||\leqslant d_{C,\eta}\}$ by Corollary A.1 and exists uniquely with probability 1 because of (10) and (12) . Thus $||\beta_P(\boldsymbol{\beta},\sigma^2)||\leqslant d_{C,\eta}\ \forall(\boldsymbol{\beta},\sigma^2)\in C$.

Now, $P^\infty$-a.s., for arbitrary $\kappa>0$:

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C} |P_nf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2} - Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}|\to 0,$$

$$\sup_{(\boldsymbol{\beta},\sigma^2,t)\in W(C,\kappa)} |P_nf_{\mathbf{t},\boldsymbol{\beta},\sigma^2} - Pf_{\mathbf{t},\boldsymbol{\beta},\sigma^2}|\to 0,$$

where

$$W(C,\kappa) := \{(\boldsymbol{\beta},\sigma^2)\in C,\ ||\mathbf{t}-\beta_P(\boldsymbol{\beta},\sigma^2)||>\kappa\}\cap\{||\mathbf{t}||\leqslant d_{C,\eta}\},$$

thus, by Proposition A.5,

$$P^\infty\left\{\exists n_0\ \forall n\geqslant n_0: \sup_{(\boldsymbol{\beta},\sigma^2)\in C}||\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2)-\beta_P(\boldsymbol{\beta},\sigma^2)||<\kappa\right\}=1. \tag{A.7}$$

Further, by definition,

$$\sigma^2_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma)=\frac{nP_nf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}}{(n-p-1)P_n\{w_{\boldsymbol{\beta},\sigma^2}=1\}},$$

$$\sigma^2_P(\boldsymbol{\beta},\sigma^2)=\frac{Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}}{P\{w_{\boldsymbol{\beta},\sigma^2}=1\}}.$$

Theorem A.1 yields, $P^\infty$-a.s.,

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C}|P_n\{w_{\boldsymbol{\beta},\sigma^2}=1\}-P\{w_{\boldsymbol{\beta},\sigma^2}=1\}|\to 0$$

and

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C}|P_nf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}-Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}|\to 0,$$

since, by (A.6), $P^\infty$-a.s.,

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C}|P_nf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}-Pf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}|\to 0,$$

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C}|P_nf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}-Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}|\to 0,$$

$$P_nf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}\geqslant P_nf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2},$$

$$Pf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}\geqslant Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}.$$

Observe, $P^\infty$-a.s.,

$$\sup_{(\boldsymbol{\beta},\sigma^2)\in C}\left|\frac{nP_nf_{\beta_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}}{(n-p-1)P_n\{w_{\boldsymbol{\beta},\sigma^2}=1\}}-\frac{Pf_{\beta_P(\boldsymbol{\beta},\sigma^2),\boldsymbol{\beta},\sigma^2}}{P\{w_{\boldsymbol{\beta},\sigma^2}=1\}}\right|\to 0,$$

since the denominators are guaranteed to be $P^\infty$-a.s. non-zero for large enough $n$ by (12). This proves Lemma 5.1 together with (A.7). □

**Proof of Theorem 5.1.** Because of Theorem 3.1 and (10), there exist LS-FPCV $\mathbf{w}_n$ w.r.t. $\mathbf{Z}_n$ for all $n>p+1$ with probability 1. For sufficiently large $n$, there exist $P^\infty$-a.s. all $f_{\mathbf{Z}_n}(w_{\boldsymbol{\beta},\sigma^2})$, $(\boldsymbol{\beta},\sigma^2)\in B_{\varepsilon_1}(\boldsymbol{\beta}_0,\sigma_0^2)$ for some $\varepsilon_1>0$ because it can be shown that

$$\exists\varepsilon_1>0: P\left(\bigcap_{(\boldsymbol{\beta},\sigma^2)\in B_{\varepsilon_1}(\boldsymbol{\beta}_0,\sigma_0^2)}\{w_{\boldsymbol{\beta},\sigma^2}=1\}\right)>0. \tag{A.8}$$

Choose $\varepsilon,\kappa>0$ small enough that $\alpha\varepsilon+\kappa<\varepsilon<\min(\varepsilon_0,\varepsilon_1)$. Eq. (12) follows from (A.8). Hence Lemma 5.1 can be applied. With help of (14) get

$$P^\infty\{\exists n_0>p+1\ \forall n>n_0,(\boldsymbol{\beta},\sigma^2)\in B_\varepsilon(\boldsymbol{\beta}_0,\sigma_0^2):$$
$$f_{\mathbf{Z}_n}(\boldsymbol{\beta},\sigma^2)\in B_\varepsilon(\boldsymbol{\beta}_0,\sigma_0^2)\}=1. \tag{A.9}$$

Thus, the fixed point algorithm started with $w_{\beta,\sigma^2}$, $(\beta,\sigma^2)\in B_\varepsilon(\beta_0,\sigma_0^2)$, stays inside of $B_\varepsilon(\beta_0,\sigma_0^2)$ with probability 1. It converges by Theorem 3.1 and (10), and therefore $f_{\mathbf{Z}_n}$ has a fixed point almost surely for all $n>p+1$, and for $n>n_0$ it can be found in $B_\varepsilon(\beta_0,\sigma_0^2)$. Let $(\eta_i)_{i\in\mathbb{N}}$ a sequence with $\eta_i\searrow 0$. Let

$$U := \bigcap_{\eta_i, i\in\mathbb{N}} \{\exists n_0<p+1\ \forall n>n_0\ \exists \mathbf{w}_{\mathbf{Z}_n,\beta_n,\sigma_n^2}\ \text{LS-FPCV w.r.t.}\mathbf{Z}_n:$$
$$||f_{\mathbf{Z}_n}(\beta_n,\sigma_n^2)-(\beta_0,\sigma_0^2)||<\eta_i\},$$

and observe $P^\infty(U)=1$, which proves the theorem. □

**Proof of Theorem 6.1.** Observe under $\sigma_1^2=0$

$$\sigma_P^2(\beta_1,\sigma^2)=0<\sigma_P^2(\beta,\sigma^2)\ \forall\sigma^2,\quad \beta\neq\beta_1,$$

i.e., $(\beta_1,0)$ is the only fixed point of $f_P$. In the following $\sigma_1^2>0$. Because of Remark 1, assume w.l.o.g. $\beta_1=\mathbf{0},\sigma_1^2=1$, i.e., $\mathbf{x}$ and $y$ are stochastically independent under $P$. The proof proceeds as follows:

*Step* 1: $h$ has a unique zero.

*Step* 2: $w_{0,\sigma^2}$ is FPCI w.r.t. $P$ iff $\sigma^2=k$ (straightforward).

*Step* 3: If $\beta\neq\mathbf{0}$, $w_{\beta,\sigma^2}$ is not FPCI w.r.t. $P$. □

**Proof of Step 1.** Use

$$h(s^2)=0 \Leftrightarrow h_0(s)$$
$$:=(1-s^2)[\Phi(\sqrt{c}s)-\Phi(-\sqrt{c}s)]-2\sqrt{c}s\varphi(\sqrt{c}s)=0.$$

Observe $s\geqslant 1\Rightarrow h_0(s)<0$, $h_0(0)=0$, and show $h_0(s)>0$ in some neighborhood of 0:

$$h_0'(s)=2s[\sqrt{c}s(c-1)\varphi(\sqrt{c}s)-(\Phi(\sqrt{c}s)-\Phi(-\sqrt{c}s))]$$
$$>2\sqrt{c}s^2[(c-1)\varphi(\sqrt{c}s)-2\varphi(0)]>0$$

in some neighborhood of 0 since $c>3$, $(c-1)\varphi(\sqrt{c}s)-2\varphi(0)>0$. Continuity of $h_0$ ensures the existence of some zero argument $>0$.

To show uniqueness of this zero, use

$$h_0'(s)=0 \Leftrightarrow h_1(s):=\sqrt{c}s(c-1)\varphi(\sqrt{c}s)-(\Phi(\sqrt{c}s)-\Phi(-\sqrt{c}s))=0,$$
$$\lim_{s\to\infty} h_1(s)=-1.$$

Notice $h_1(0)=0$. $h_1$ has the same sign as $h_0'$ for all positive arguments. Calculate

$$h_1'(s)=\sqrt{c}\varphi(\sqrt{c}s)[(c-1)(1-cs^2)-2],$$
$$h_1'(0)=\sqrt{c}\varphi(0)(c-3)>0.$$

$h_1'(s)<0$ iff $0>(c-1)(1-cs^2)-2<0$, which is strictly monotone decreasing in $s^2$. Thus, $h_1'$ has a unique zero $s_2$, which is a local maximum of $h_1$, $h_1(s_2)>0$. $h_1$ decreases strictly monotonously for $s>s_2$ and must have a unique zero, which is the unique local extremum of $h_0$. Thus, $h_0$ can only have a unique zero. □

**Proof of Step 3.** Suppose that $w_{\boldsymbol{\beta},\sigma^2}$ with $\boldsymbol{\beta}\neq\mathbf{0}$ is FPCI w.r.t. $P$. $\sigma^2=0$ is impossible since $P\{w_{\boldsymbol{\beta},0}=1\}=0$. Define

$$F_{\boldsymbol{\beta}}(\mathbf{t}) := P(y-\mathbf{x}'\mathbf{t})^2 1((y-\mathbf{x}'\boldsymbol{\beta})^2 \leqslant cs^2),$$

i.e., $\beta_P(\boldsymbol{\beta},\sigma^2)=\arg\min_{\mathbf{t}} F_{\boldsymbol{\beta}}(\mathbf{t})$. With $\mathbf{v} := \frac{\boldsymbol{\beta}}{||\boldsymbol{\beta}||}$ get

$$\frac{\partial}{\partial \mathbf{v}} F_{\boldsymbol{\beta}}(\mathbf{t}) = -2\sum_{i=1}^{p} v_i P\mathbf{x}_i (y-\mathbf{x}'\mathbf{t}) 1((y-\mathbf{x}'\boldsymbol{\beta})^2 \leqslant c\sigma^2),$$
$$\frac{\partial}{\partial \mathbf{v}} F_{\boldsymbol{\beta}}(\boldsymbol{\beta}) = -\frac{2}{||\boldsymbol{\beta}||} G[\mathbf{x}'\boldsymbol{\beta} J(\mathbf{x}'\boldsymbol{\beta})],$$

where

$$J(u) := \mathcal{N}(y-u)1((y-u)^2 \leqslant c_0^2), \quad c_0 := \sqrt{cs^2}.$$

Show $uJ(u)<0$ for $u\neq 0$:

$$\begin{aligned} uJ(u) &= \int u(y-u)1[|y-u|\leqslant c_0]\varphi(y)\,dy \\ &= \int u|y-u|1[|y-u|\leqslant c_0](1[y>u]-1[y<u])\varphi(y)\,dy \\ &= \int u|t|1[0<t\leqslant c_0](\varphi(t+u)-\varphi(-t+u))\,dt \\ &= \int |u|\,|t|1[0<t\leqslant c_0](\varphi(t+|u|)-\varphi(-t+|u|))\,dt, \end{aligned}$$

since

$$\varphi(t+u)=\varphi(-t+|u|), \quad \varphi(-t+u)=\varphi(t+|u|)$$

for $u<0$. Get $uJ(u)<0$ by

$$t>0, w>0 \Rightarrow wt[\varphi(t+w)-\varphi(-t+w)]<0.$$

Since $G\mathbf{x}\mathbf{x}'$ was supposed to be invertible in (6), $G\{\mathbf{x}'\boldsymbol{\beta}=0\}<1$. That is, $\frac{\partial}{\partial \mathbf{v}} F_{\boldsymbol{\beta}}(\boldsymbol{\beta})>0$, and $\boldsymbol{\beta}\neq\beta_P(\boldsymbol{\beta},\sigma^2)$. The proof is completed. □

### *A.3. Preparations for the proof of Lemma 6.1*

**Theorem A.2.** *Let $K$ be a compact convex subset of $\mathbb{R}^p$, $C^1(K)\ni f=(f_1,\ldots,f_q): K\mapsto\mathbb{R}^q$. Then,*

$$\forall \mathbf{x},\mathbf{y}\in K: ||f(\mathbf{x})-f(\mathbf{y})|| \leqslant ||df||_K ||\mathbf{x}-\mathbf{y}||,$$

*where*

$$||\mathrm{d}f||_K := \sup_{\mathbf{x}\in K}\left(\max_{i=1,\ldots,q}\sum_{j=1}^{p}\left|\left[\frac{\partial}{\partial x_j}f_i\right](\mathbf{x})\right|\right).$$

Proof e.g. in [23].

**Proposition A.6.** *Let* $h_1:\mathbb{R}^{p+1}\mapsto\mathbb{R}$ *be continuous, where* $h_1$, $h_{1i}(\mathbf{x}) := x_i h_1(\mathbf{x})$ *G-integrable for* $i=1,\ldots,p+1$, $h_2:\mathbb{R}\mapsto\mathbb{R}$ *be continuous,* $\mathcal{N}$*-integrable and* $\exists y_0<\infty:|h_2\varphi|\leqslant y_0$. *Then,* $l:\mathbb{R}^{p+2}\mapsto\mathbb{R}$ *defined by*

$$l(a,\mathbf{b}) := \int h_1(\mathbf{x})h_2(y)1[(y-\mathbf{x}'\mathbf{b})^2\leqslant a^2]\varphi(y)d[\lambda(y)\otimes G(\mathbf{x})]$$

*is continuously differentiable. The proof is straightforward.*

**Proof of Lemma 6.1.** Because of the Remarks 1 and 2, assume w.l.o.g. $\boldsymbol{\beta}_1=0,\sigma_1^2=1$, i.e., $\mathbf{x}$ and $y$ are stochastically independent under $P_{\mathbf{0},1,G}$. Consider (18) and $\mathcal{L}(y)=\mathcal{N}_{(0,1)}$ under $P_{\mathbf{0},1,G}$, and get (9) by continuity of $\mathcal{L}(y)$, (10) and (11) by (17). $w_{\mathbf{0},k}$ is LS-FPCI w.r.t. $P$ and $k$ is uniquely defined by Corollary 2, which requires $P^*\{y^2\leqslant ck\}=0$ by (18), thus (13) is fulfilled. Show (15) by application of Theorem A.2 and Proposition A.6:

To prove (15), it suffices to have $||df_P||_{B_{\varepsilon_0}(\mathbf{0},k)}=:\alpha<1$ for some $\varepsilon_0>0$. It can be shown by use of Proposition A.6 that $f_P$ is continuously differentiable in $(\mathbf{0},k)$. The proof is completed by showing

$$\left(\max_{i=1,\ldots,p+2}\sum_{j=1}^{p+2}\left|\left[\frac{\partial}{\partial x_j}f_{Pi}\right](\mathbf{0},k)\right|\right)<1. \tag{A.10}$$

This can be done by computing the partial derivatives of $f_{P_{\mathbf{0},1,G}}$ in $(\mathbf{0},k)$. $\beta_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2)$ is defined by

$$F(\boldsymbol{\beta},\sigma^2,\beta_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2))=0,$$

where

$$F(\boldsymbol{\beta},\sigma^2,\mathbf{t}) := (P_{\mathbf{0},1,G}\mathbf{x}\mathbf{x}'1[(y-\mathbf{x}'\boldsymbol{\beta})^2\leqslant c\sigma^2])\mathbf{t}-P_{\mathbf{0},1,G}\mathbf{x}y1[(y-\mathbf{x}'\boldsymbol{\beta})^2\leqslant c\sigma^2].$$

By Proposition A.6, $F$ is continuously differentiable w.r.t. $\boldsymbol{\beta},\sigma^2,\mathbf{t}$. If $\sigma^2>0$, then $P_{\mathbf{0},1,G}\{w_{\boldsymbol{\beta},\sigma^2}=1\}>0$ for arbitrary $\boldsymbol{\beta}$. $\frac{\partial}{\partial\mathbf{t}}F(\boldsymbol{\beta},\sigma^2,\mathbf{t})=P_{\mathbf{0},1,G}\mathbf{x}\mathbf{x}'1[(y-\mathbf{x}'\boldsymbol{\beta})^2\leqslant c\sigma^2]$ is invertible by (17), $\beta_{P_{\mathbf{0},1,G}}$ is continuous at $\boldsymbol{\beta},\sigma^2$ (Proposition A.2). Notice $\beta_{P_{\mathbf{0},1,G}}(\mathbf{0},k)=0$ by step 2 of the proof of Theorem 6.1. Differentiation of implicit functions

(see [20]) yields

$$\left.\frac{\partial F(\boldsymbol{\beta},\sigma^2,\mathbf{t})}{\partial \mathbf{t}}\right|_{(\boldsymbol{\beta},\sigma^2,\mathbf{t})=(\mathbf{0},k,\mathbf{0})} = [\Phi(\sqrt{ck})-\Phi(\sqrt{-ck})]G\mathbf{x}\mathbf{x}',$$

$$\left.\frac{\partial F(\boldsymbol{\beta},\sigma^2,\mathbf{t})}{\partial \sigma^2}\right|_{(\boldsymbol{\beta},\sigma^2,\mathbf{t})=(\mathbf{0},k,\mathbf{0})} = 0,$$

$$\left.\frac{\partial F(\boldsymbol{\beta},\sigma^2,\mathbf{t})}{\partial \boldsymbol{\beta}}\right|_{(\boldsymbol{\beta},\sigma^2,\mathbf{t})=(\mathbf{0},k,\mathbf{0})} = -2\sqrt{ck}\varphi(\sqrt{ck})G\mathbf{x}\mathbf{x}'$$

$$\Rightarrow \left.\frac{\partial}{\partial \boldsymbol{\beta}}\beta_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2)\right|_{(\boldsymbol{\beta},\sigma^2)=(\mathbf{0},k)} = \frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck})-\Phi(\sqrt{-ck})}\mathbf{I}_{p+1},$$

$$\left.\frac{\partial}{\partial \sigma^2}\beta_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2)\right|_{(\boldsymbol{\beta},\sigma^2)=(\mathbf{0},k)} = 0.$$

Notice $1>1-E_{\mathcal{N}}(y^2|y^2\leqslant ck)=\frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck})-\Phi(\sqrt{-ck})}>0$.

Now evaluate

$$\frac{\partial}{\partial \sigma^2}\sigma^2_{P_{\mathbf{0},1,G}}(\mathbf{0},k) = (c-1)\frac{1-k}{2},$$

get $|\frac{\partial}{\partial \sigma^2}\sigma^2_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2)|_{(\boldsymbol{\beta},\sigma^2)=(\mathbf{0},k)}|<1$ by (19), and observe by symmetry considerations $\sigma^2_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2)=\sigma^2_{P_{\mathbf{0},1,G}}(-\boldsymbol{\beta},\sigma^2)$, thus $\frac{\partial}{\partial \boldsymbol{\beta}}\sigma^2_{P_{\mathbf{0},1,G}}(\boldsymbol{\beta},\sigma^2)|_{(\boldsymbol{\beta},\sigma^2)=(\mathbf{0},k)}=0$.

Altogether,

$$||df_P||_{B_{\varepsilon_0}(\mathbf{0},k)} \leqslant \max\left(\frac{2\sqrt{ck}\varphi(\sqrt{ck})}{\Phi(\sqrt{ck})-\Phi(-\sqrt{ck})}, (c-1)\frac{1-k}{2}\right)<1,$$

proving (A.10).

This completes the proof. □

## References


[1] J.G. Adrover, V.J. Yohai, Simultaneous redescending M-estimates for regression and scale, Comm. Statist. (Theory and Methods) 29 (2000) 243–262.

[2] A. Azzalini, A.W. Bowman, A look at some data on the Old Faithful geyser, J. Roy. Statist. Soc. Ser. C (Appl. Statist.) 39 (1990) 357–365.

[3] S. Byers, A.E. Raftery, Nearest neighbor clutter removal for estimating features in spatial point processes, J. Amer. Statist. Assoc. 93 (1998) 577–584.

[4] D. Comaniciu, P. Meer, Mean shift analysis and applications, IEEE International Conference on Computer Vision, Keskyra, Greece, 1999, pp. 1197–1203.

[5] R.D. Cook, F. Critchley, Identifying regression outliers and mixtures graphically, J. Amer. Statist. Assoc. 95 (2000) 781–794.

[6] R.D. Cook, S. Weisberg, Residuals and Influence in Regression, Chapman & Hall, London, 1982.

[7] T.H. Cormen, C.E. Leiserson, R.L. Rivest, Introduction to Algorithms, MIT Press, Cambridge, 1990.

[8] J.A. Cuesta-Albertos, A. Gordaliza, C. Matran, Trimmed $k$-means: an attempt to robustify quantizers, Ann. Statist. 25 (1997) 553–576.

[9] A. DasGupta, A.E. Raftery, Detecting features in spatial point processes with clutter via model-based clustering, J. Amer. Statist. Assoc. 93 (1998) 294–302.
[10] P.L. Davies, Consistent estimates for finite mixtures of well separated elliptical distributions, in: H.-H. Bock (Ed.), Classification and Related Methods of Data Analysis, Elsevier Science Publishers, Amsterdam, 1988.
[11] P.L. Davies, U. Gather, The identification of multiple outliers, with discussion, J. Amer. Statist. Assoc. 88 (1993) 782–801.
[12] W.S. DeSarbo, W.L. Cron, A maximum likelihood methodology for clusterwise linear regression, J. Classification 5 (1988) 249–282.
[13] R.C. Fair, D.M. Jaffee, Methods of estimation for markets in disequilibrium, Econometrica 40 (1972) 497–514.
[14] C. Fraley, A.E. Raftery, How many clusters? Which clustering method? Answers via model based cluster analysis, Comput. J. 41 (1998) 578–588.
[15] L.A. Garcia-Escudero, A. Gordaliza, Robustness properties of $k$ means and trimmed $k$ means, J. Amer. Statist. Assoc. 94 (1999) 956–969.
[16] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. Stahel, Robust Statistics, The Approach Based on Influence Functions, Wiley, New York, 1986.
[17] C. Hennig, N. Christlieb, Validating visual clusters in large dataset fixed point clusters of spectral features, Comput. Statist. Data Anal. 40 (2002) 723–739.
[18] C. Hennig, What clusters are generated by normal mixtures?, in: H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, M. Schaden (Eds.), Data Analysis, Classification and Related Methods, Springer, Berlin, 2000a, pp. 53–58.
[19] C. Hennig, Identifiability of models for clusterwise linear regression, J. Classification 17 (2000b) 273–296.
[20] C. Hennig, Regression fixed point clusters: motivation, consistency and simulations, preprint 2000-02, 2000c, Fachbereich Mathematik—SPST, University of Hamburg, http://www.math.uni-hamburg.de/home/hennig|.
[21] D.W. Hosmer Jr., Maximum likelihood estimates of the parameters of a mixture of two regression lines, Comm. Statist. 3 (1974) 995–1006.
[22] P.J. Huber, Robust Statistics, Wiley, New York, 1981.
[23] K. Königsberger, Analysis 2, Springer, Berlin, 1993.
[24] G. McLachlan, D. Peel, Finite Mixture Models, Wiley, New York, 2000.
[25] S. Morgenthaler, Fitting redescending M-estimators in regression, in: H.D. Lawrence, S. Arthur (Eds.), Robust Regression, Dekker, New York, 1990, pp. 105–128.
[26] P.J. Rousseeuw, M. Hubert, Recent developments in PROGRESS, in: Y. Dodge (Ed.), L1-Statistical Procedures and Related Topics, IMS Lecture Notes, Vol. 31, IMS Lecture Notes and Monograph Series, Berkeley, USA, 1997, pp. 201–214.
[27] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, Wiley, New York, 1987.
[28] A.W. van der Vaart, J.A. Wellner, Weak Convergence and Empirical Processes, Springer, New York, 1996.
[29] M. Wedel, W.S. DeSarbo, A mixture likelihood approach for generalized linear models, J. Classification 12 (1995) 21–56.
[30] C. Hennig, Fixed point clusters for linear regression: computation and comparison, J. Classification 19 (2002) 249–276.